Four Challenges for Statistical Agencies

John M. Abowd
Cornell University and U.S Census Bureau*
BLS Commissioner's Invited Seminar
May 16, 2016

*Prepared as a Cornell professor before I assumed an executive role at the Census Bureau. My Census Bureau appointment begins on June 1, 2016.

I want to begin by thanking the Commissioner for this invitation. As a professional economist and statistician for 40 years, I have always regarded the Bureau of Labor Statistics with the highest esteem as a model producer of quality information in an admirably transparent and professional manner. Your *Handbook of Methods* is the international standard for organizing the metadata for a statistical agency's products.

There is no question in my mind that you understand what you are doing, and how important that work is for keeping the world informed about social and economic conditions in the United States.

I don't mean to be presumptuous in presenting these challenges, many of which you have already recognized and begun to address. Still, we have a long way to travel, and I hope that my remarks will be helpful in navigating that journey.

The Data Bloggers

INTEREST IN SELF-INDUCED ABORTION

Google search rate above or below national average for phrases like “home abortion methods,” 2011 to 2015.

ABOVE AVERAGE

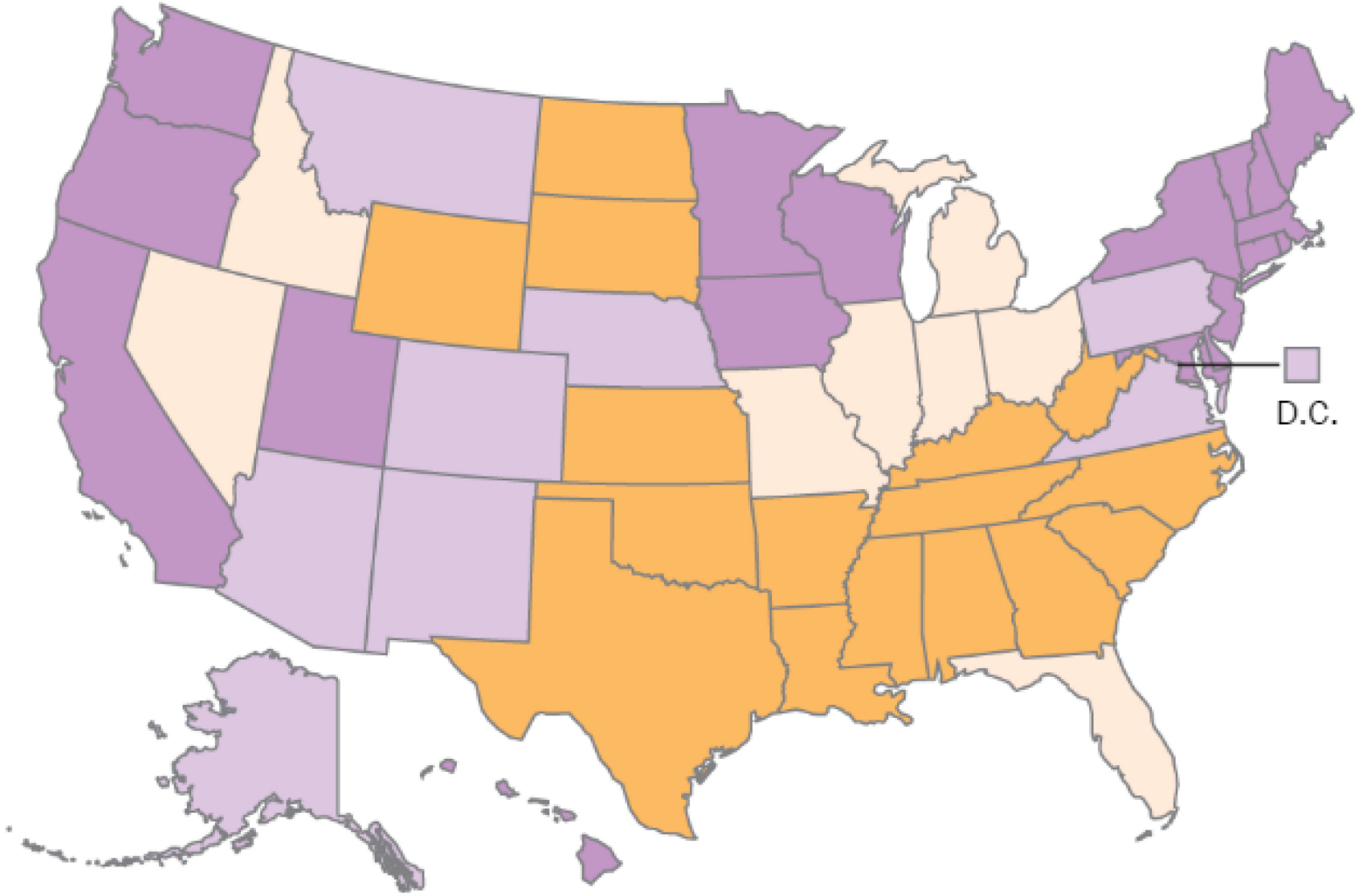
More than 10% above

Less than 10%

BELOW AVERAGE

Less than 10% below

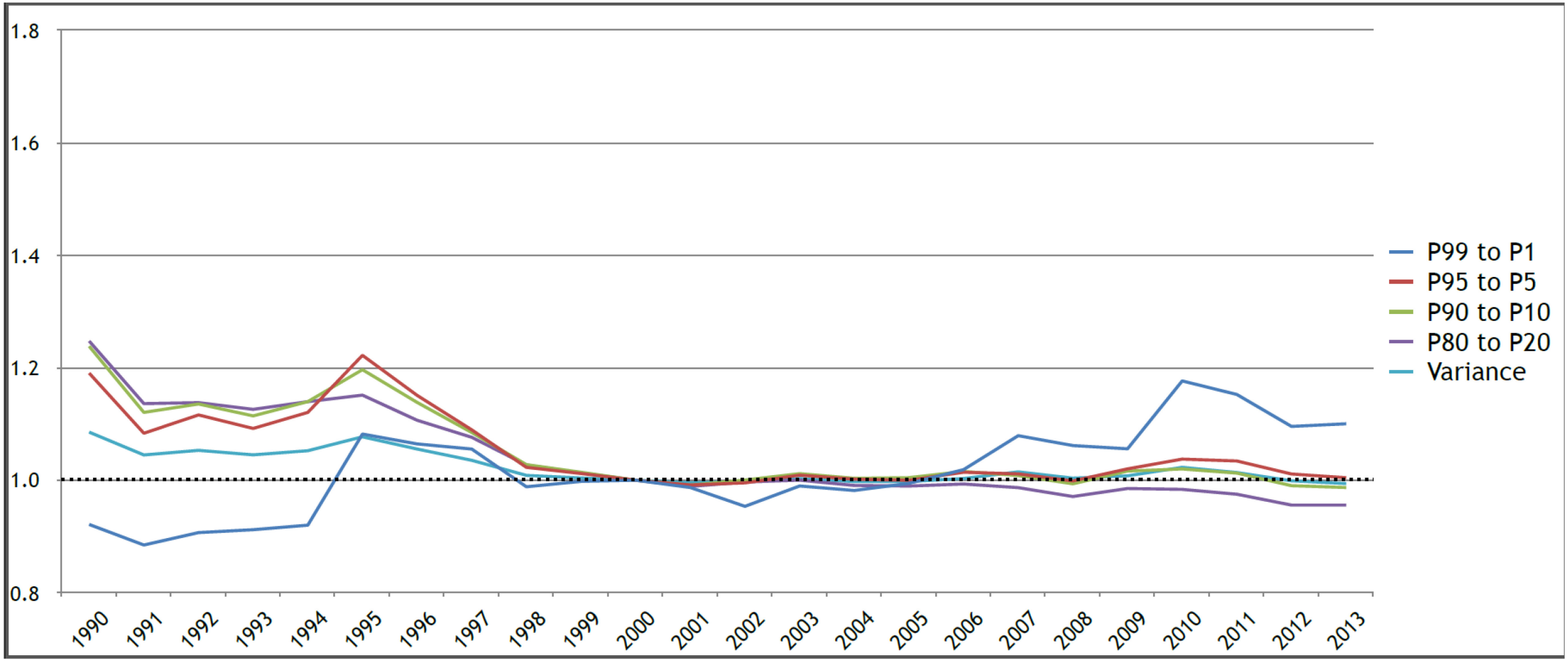
More than 10%



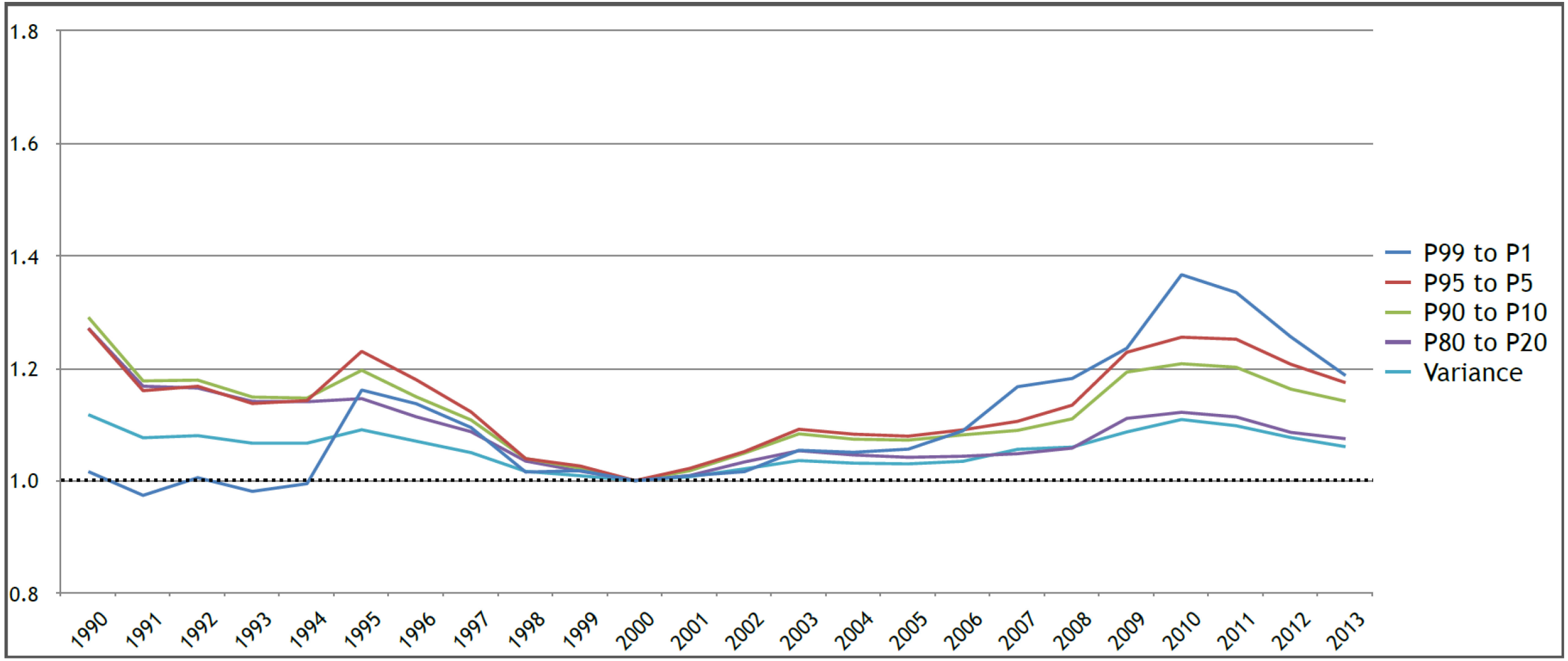
Sources: Guttmacher Institute (state laws); analysis of Google data by Seth Stephens-Davidowitz (searches)

By Bill Marsh/The New York Times

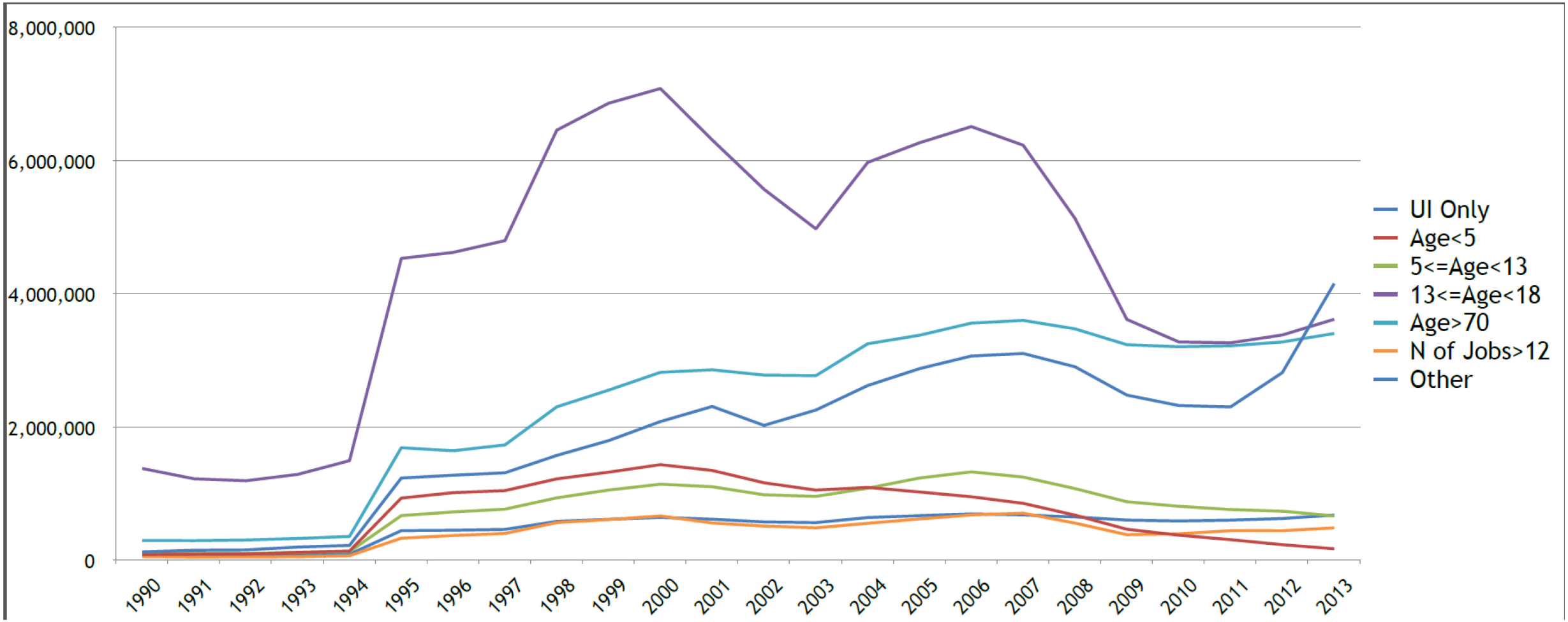
Earnings Inequality in the U.S.-Uncorrected



Earnings Inequality in the U.S.-Corrected

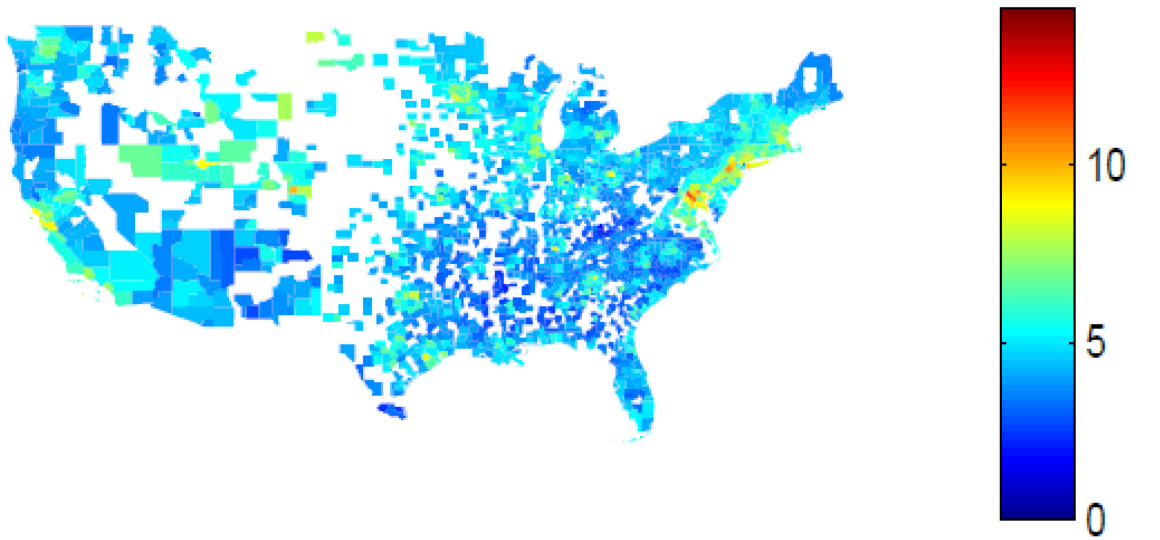


Records Removed from the Frame and Why

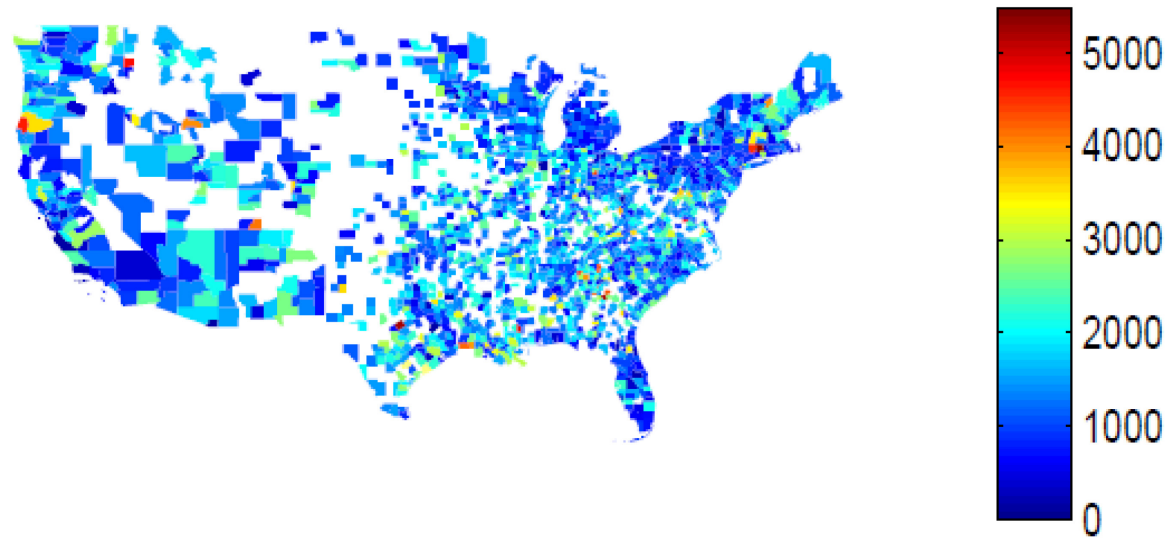


Let Me Model That for You

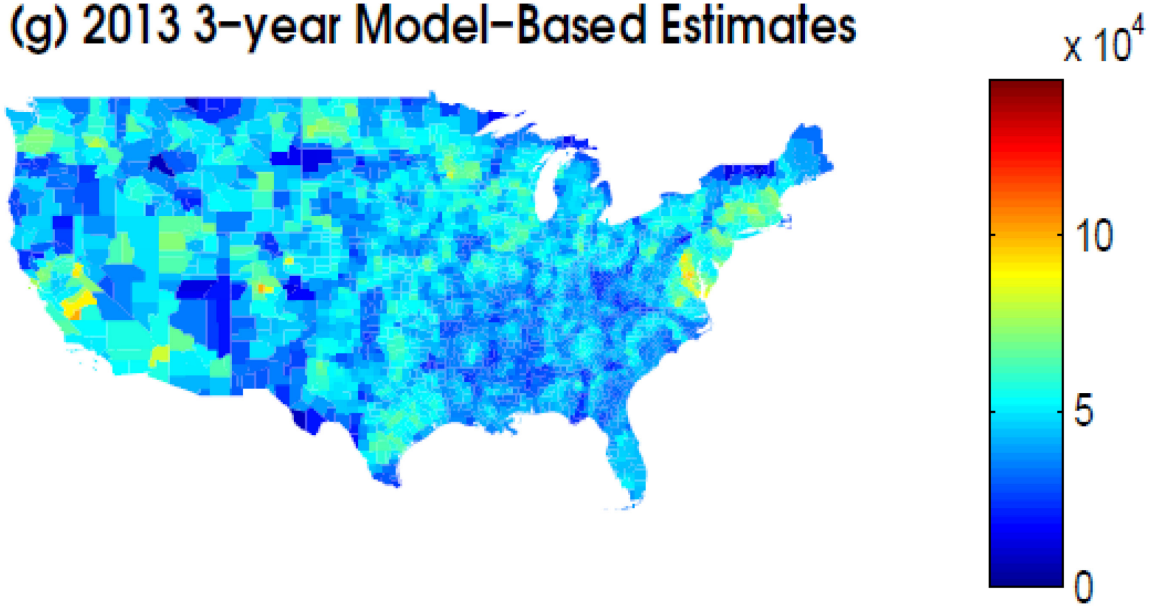
(c) 2013 3-year ACS Estimates



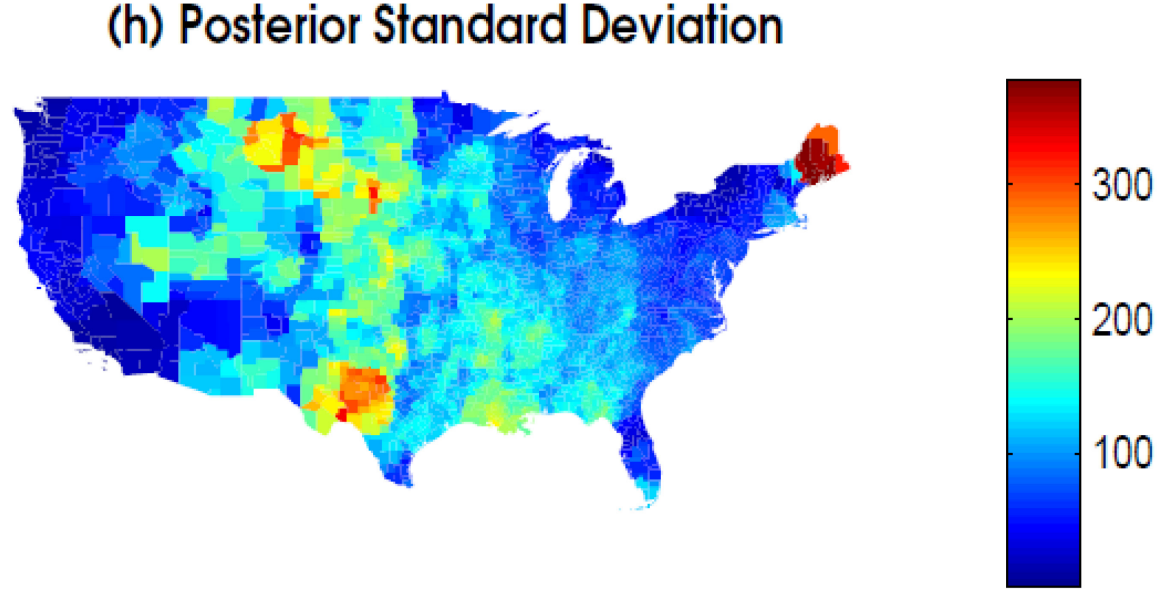
(d) 2013 3-year ACS Estimates of Std.Dev



(g) 2013 3-year Model-Based Estimates



(h) Posterior Standard Deviation



MANHATTAN

1,2 WALL STREET, CIVIC CENTER, GOVERNORS ISLAND, LIBERTY ISLAND, ELLIS ISLAND, TRIBECA, GREENWICH VILLAGE, NOHO, SOHO, LITTLE ITALY

- 42 percent more households with children
- 54 percent more people age 55 to 64
- 137 percent more residents who work in protective services (police, security, etc.)

3 LOWER EAST SIDE, CHINATOWN

- 55 percent more adults with bachelor's degrees but no higher degrees
- 43 percent more households of men living alone
- 24 percent fewer Hispanic residents

4,5 CHELSEA, HELL'S KITCHEN, HERALD SQUARE, MIDTOWN, TIMES SQUARE

- 30 percent more residents age 35 to 44
- 21 percent fewer households headed by women
- 52 percent fewer homes owned and occupied by Hispanics

6 MURRAY HILL, EAST MIDTOWN, STUYVESANT TOWN

- 33 percent fewer adults with some college education but no four-year degrees
- 42 percent fewer residents who work in transportation
- 15 percent fewer residents who are widowed, divorced or separated

7 UPPER WEST SIDE, LINCOLN SQUARE

- 34 percent fewer Hispanic families
- 24 percent more married residents
- 105 percent more children under 5

THE NEW YORK TIMES

- 46 percent fewer residents who work in construction and manufacturing

2 SUNNYSIDE, WOODSIDE

- 29 percent more residents age 55 to 64
- 39 percent fewer residents who work in construction and manufacturing
- 17 percent more residents who are widowed, divorced or separated

3 JACKSON HEIGHTS, EAST ELMHURST, NORTH CORONA

- 29 percent fewer black households

4 ELMHURST, CORONA

- 36 percent fewer blacks
- 24 percent fewer households of women living alone

5 MASPETH, RIDGEWOOD, MIDDLE VILLAGE, GLENDALE

- 56 percent more residents who work in health care
- 14 percent fewer households of women living alone
- 26 percent fewer residents with less than a high school education

6 REGO PARK, FOREST HILLS

- 47 percent more residents who work in building services (janitors, superintendents, etc.)
- 41 percent more residents who work in restaurants and food services
- 48 percent fewer black families

7 FLUSHING, WHITESTONE, COLLEGE POINT

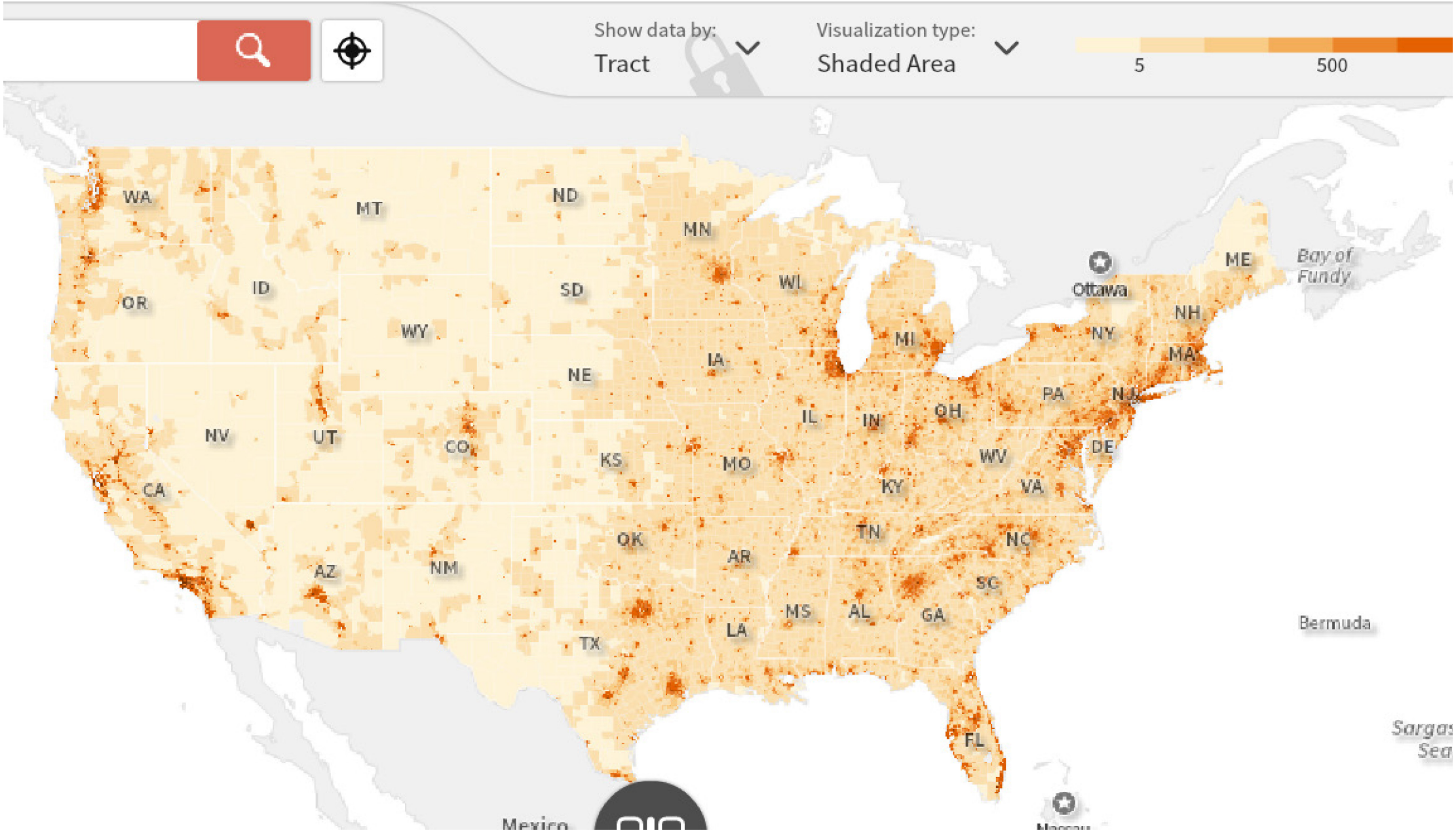
- 45 percent more employed workers
- 50 percent more residents who work in personal services

8 FRESH MEADOWS, JAMAICA HILLS, KEW GARDENS HILLS

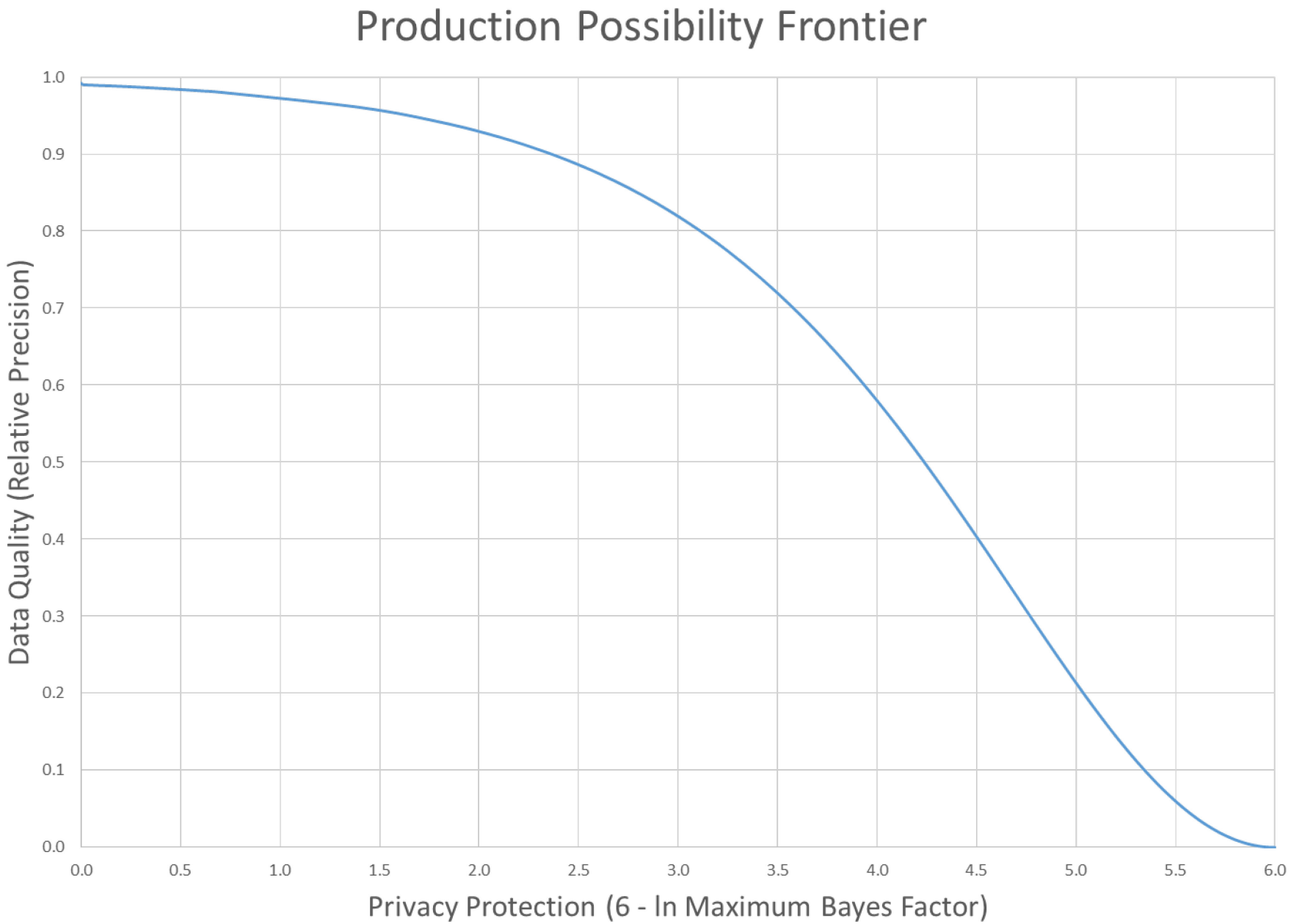
- 24 percent more households of men living alone

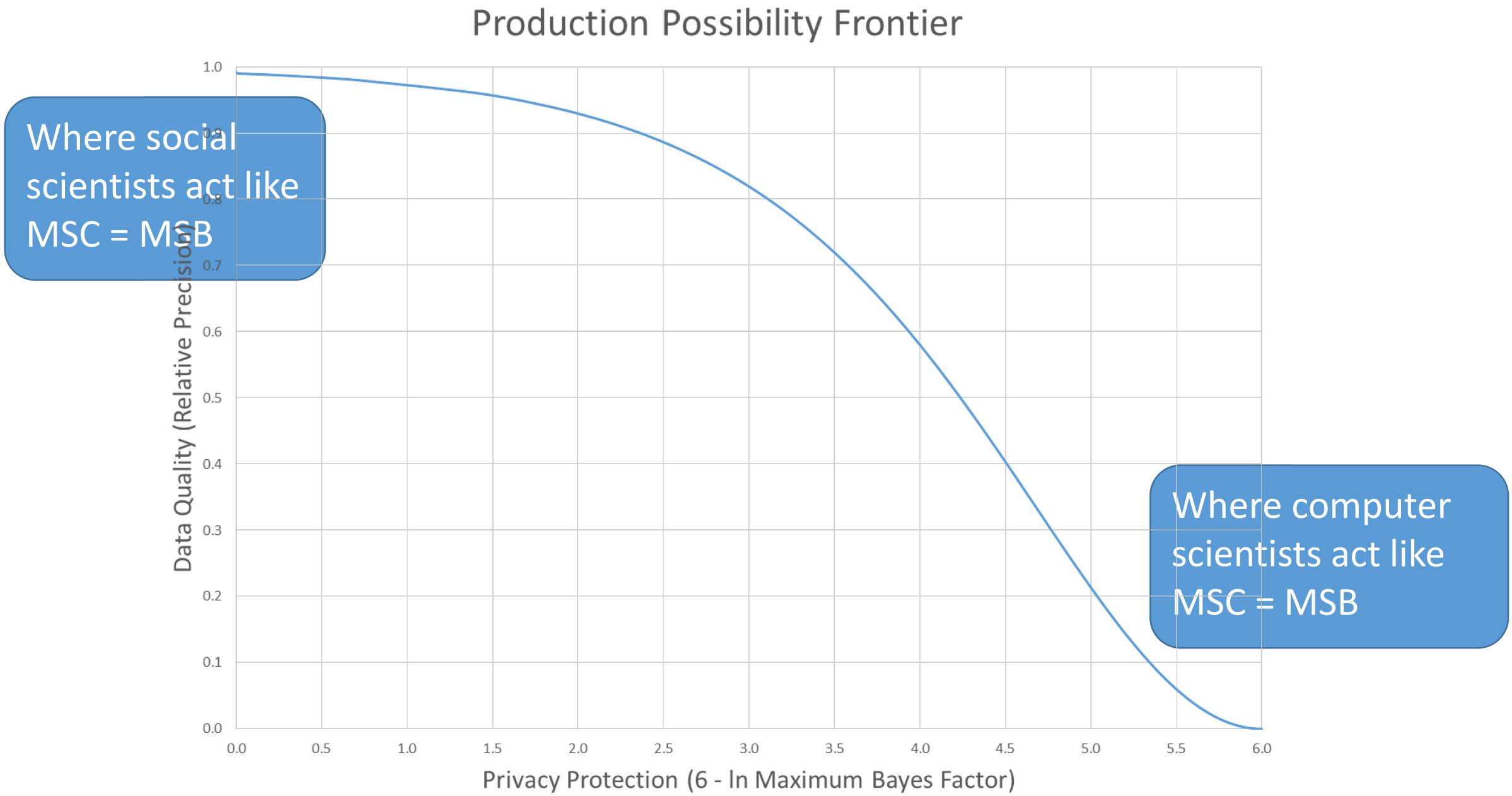
Population Density (per sq. mile)

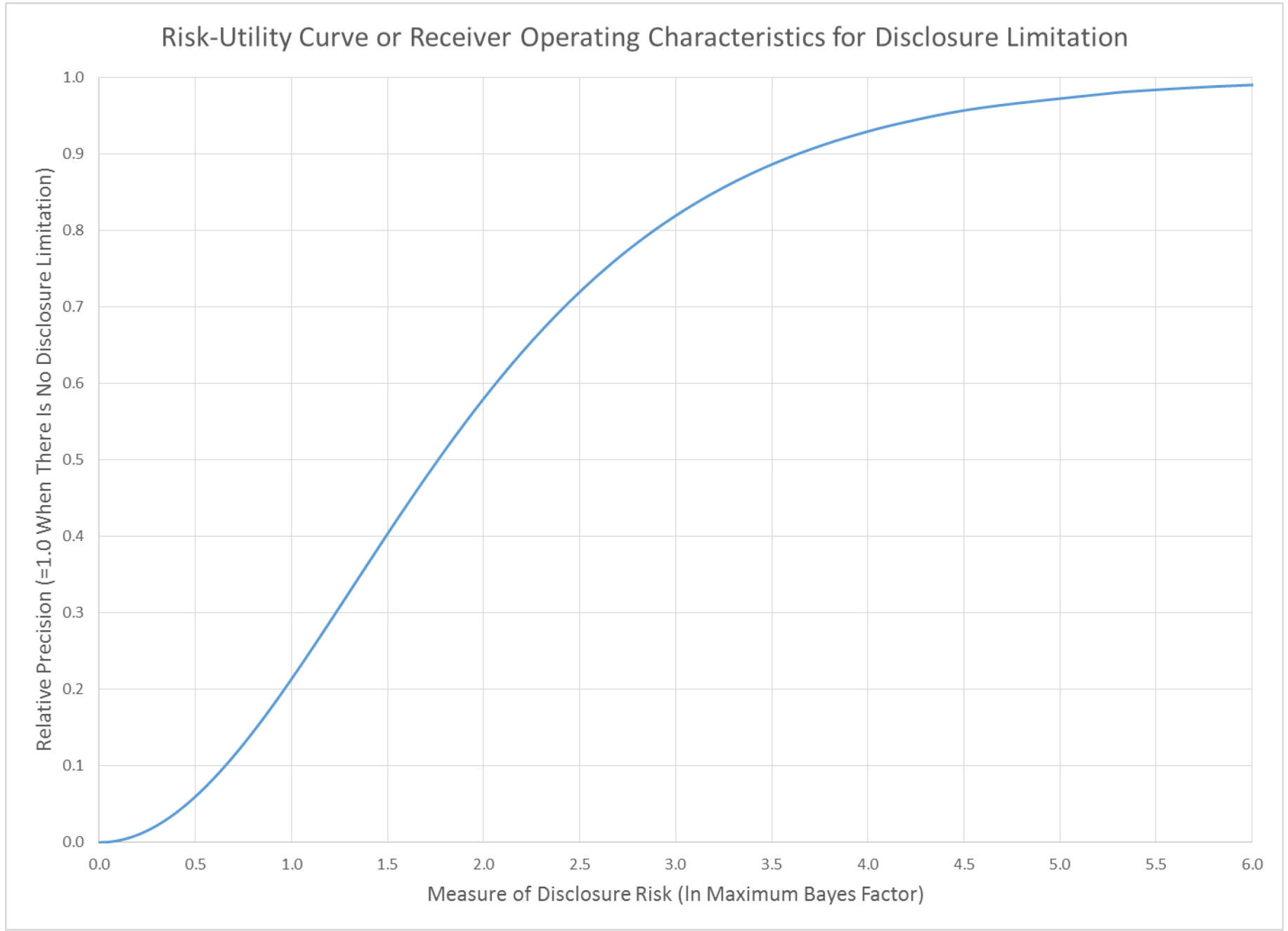
ACS 2014 (5-Year Estimates)

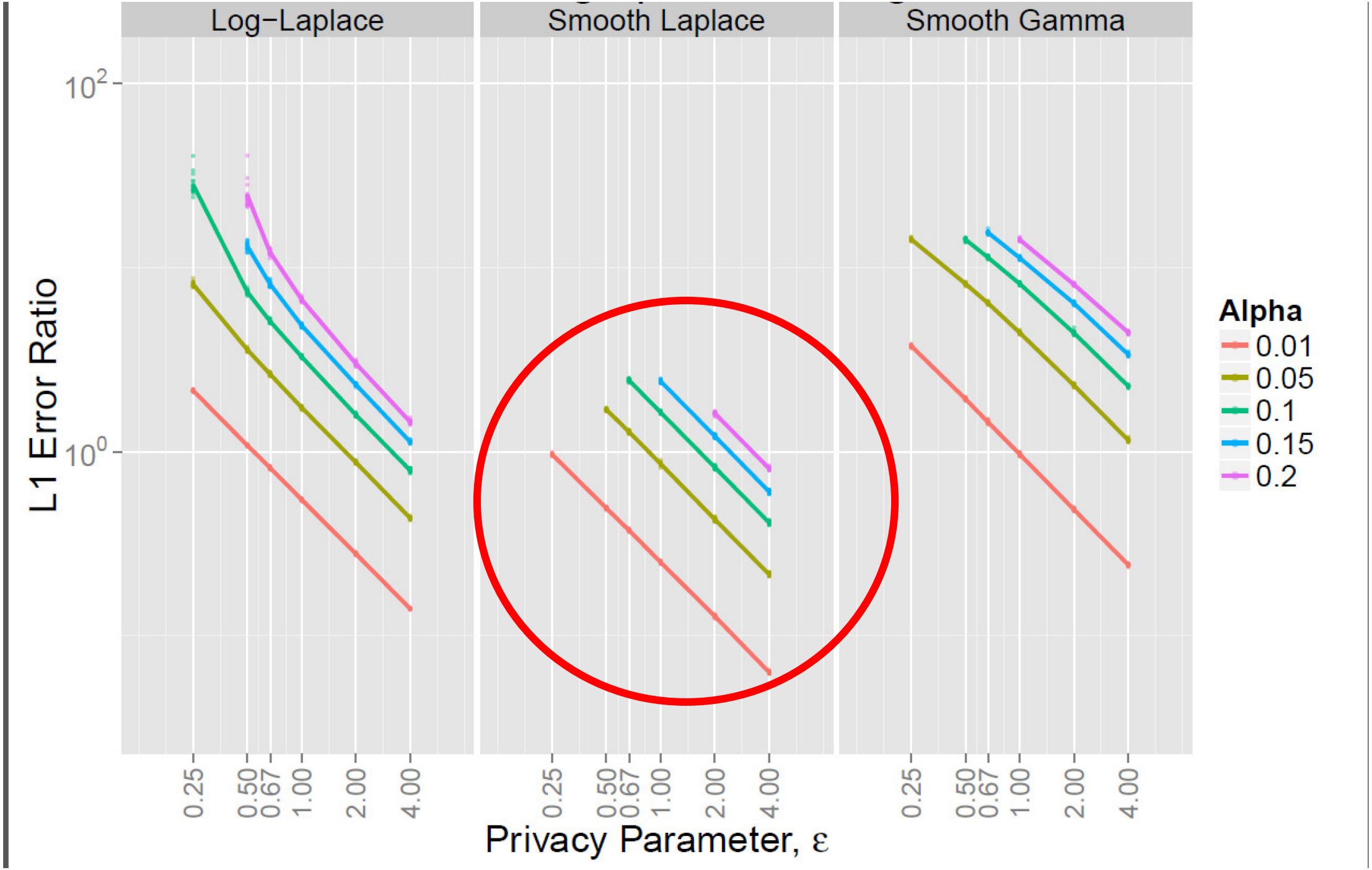


What If All Data Were Private?


Production Possibility Frontier
Data Quality (Relative Precision)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Privacy Protection (6 - ln Maximum Bayes Factor)


Production Possibility Frontier
Where social scientists act like MSC = MSB
Where computer scientists act like MSC = MSB
Data Quality (Relative Precision)
Privacy Protection (6 - ln Maximum Bayes Factor)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0


Risk-Utility Curve or Receiver Operating Characteristics for Disclosure Limitation
Relative Precision (=1.0 When There Is No Disclosure Limitation)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Measure of Disclosure Risk (In Maximum Bayes Factor)


Log−Laplace
Smooth Laplace
Smooth Gamma
L1 Error Ratio
10^2
10^0
0.25
0.50
0.67
1.00
2.00
4.00
Privacy Parameter, ε
Alpha
0.01
0.05
0.1
0.15
0.2

Can We Stop P-hacking?

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

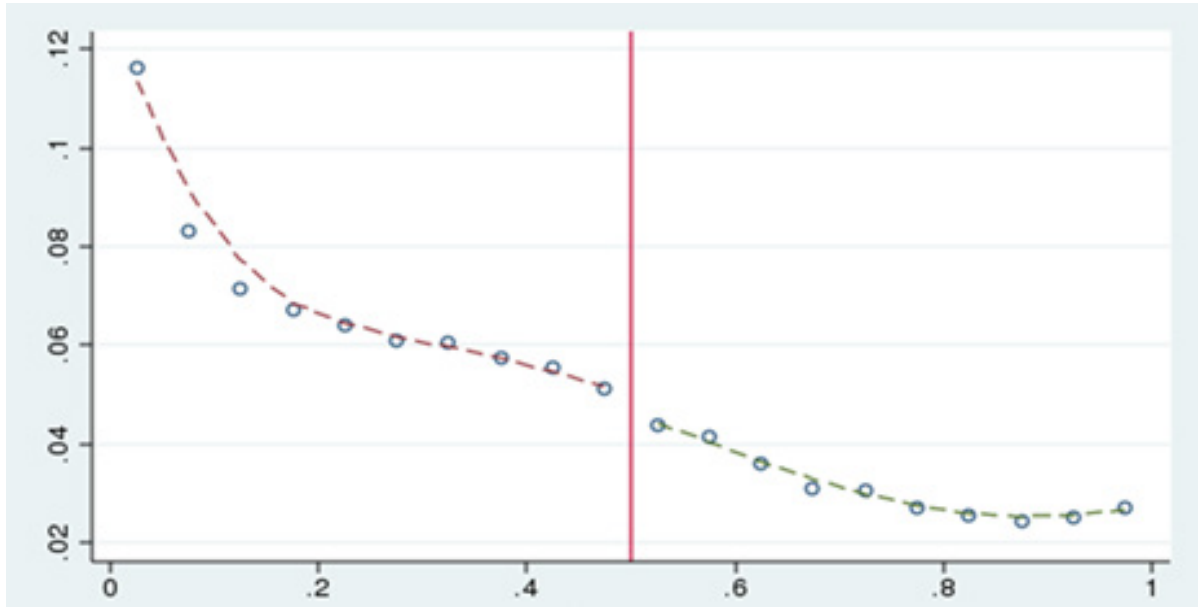
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

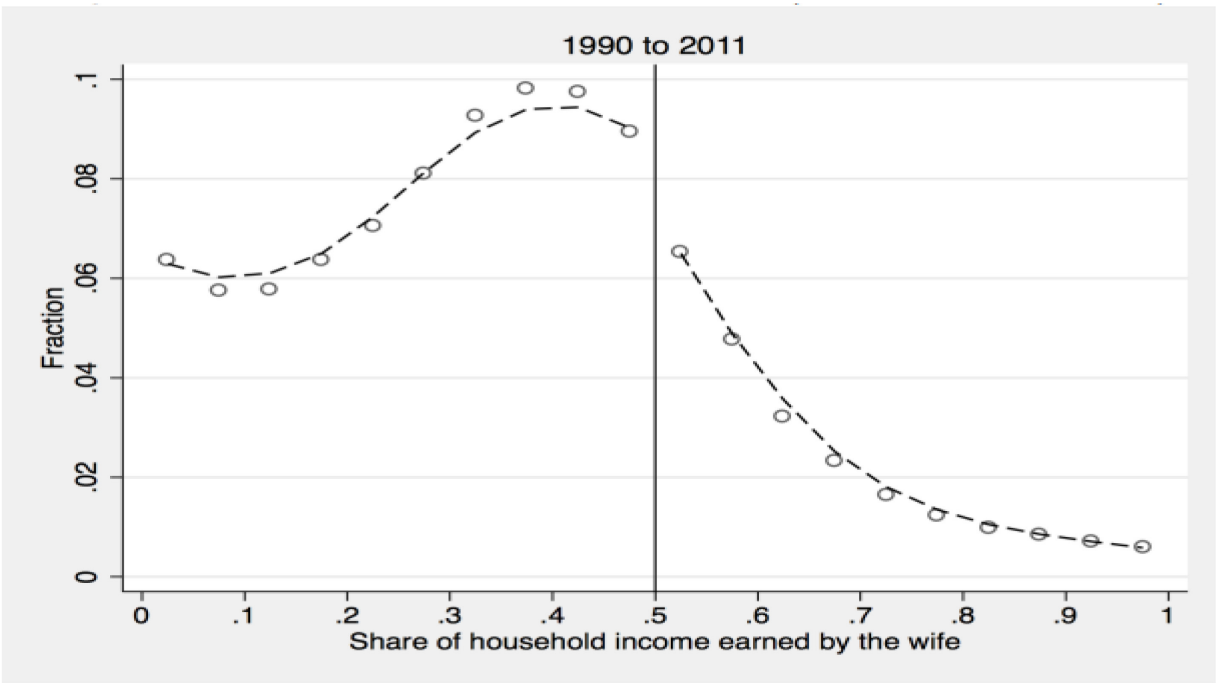
AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*





Bertrand, Kamenica and Pan (QJE 2015), doi: 10.1093/qje/qjv001

Putting the Pieces Together

Suppose we wanted to design a new, continuously updated information system on local labor markets.

Use the ideas from “Data bloggers” to properly combine contemporaneous elements harvested from the data jungle with designed elements produced by the agency.

Use the ideas from “Let me model that for you” to produce local estimates and measures of reliability for all local areas every period, including periods when the designed content is not in the field.

Use the ideas from “What if all data were private?” to provably protect the design-consistent, model-based estimates from all future privacy attacks.

Use the ideas from “Can we stop p-hacking?” to open a portal to the underlying data that returns safe estimates of hypotheses (i.e., estimates that have a controlled false discovery rate) and incorporates them into future versions of the model.

There are working prototypes of all these pieces running now. That’s where I got the graphics in this talk.

Suggested Reading

Raghunathan (2015) "Statistical Challenges in Combining Information from Big and Small Data Sources" (public version)

Stephens-Davidowitz and Varian (2015) "A Hands-on Guide to Google Data"

Bradley, Wikle and Holan (2015) "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates"

Bradley, Holan and Wikle (2015) "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics"

Abowd and Schmutte (2015) "Revisiting the Economics of Privacy"

Haney et al. (2015) "Formal privacy protection for data products combining individual and employer frames" (public version)

American Statistical Association Releases Statement on Statistical Significance and P-values

Erlingsson, Pihur and Korolova (2014 "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response"

Wang, Lei and Fienberg (2016) "On-Average KL-Privacy and its equivalence to Generalization for Max-Entropy Mechanisms"

Bertrant, Kamenica and Pan (2015) "Gender Identity and Relative Income within Households"

Abowd and Schmutte (2015) "Economic Analysis and Statistical Disclosure Limitation"

Thank you.

Contact: john.maron.abowd@census.gov